A meta-analysis on the reliability of comparative judgement

Verhavert, San; Bouwer, Renske; Donche, Vincent; De Maeyer, Sven

published in
Assessment in Education: Principles, Policy, & Practice
2019

DOI (link to publisher)
10.1080/0969594X.2019.1602027

document version
Publisher's PDF, also known as Version of record

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

A meta-analysis on the reliability of comparative judgement

San Verhavert, Renske Bouwer, Vincent Donche & Sven De Maeyer

To cite this article: San Verhavert, Renske Bouwer, Vincent Donche & Sven De Maeyer (2019) A meta-analysis on the reliability of comparative judgement, Assessment in Education: Principles, Policy & Practice, 26:5, 541-562, DOI: 10.1080/0969594X.2019.1602027

To link to this article: https://doi.org/10.1080/0969594X.2019.1602027

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
2019, VOL. 26, NO. 5, 541–562
https://doi.org/10.1080/0969594X.2019.1602027

Routledge
Taylor & Francis Group

A meta-analysis on the reliability of comparative judgement

San Verhavert, Renske Bouwer, Vincent Donche and Sven De Maeyer

Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium

ABSTRACT

Comparative Judgement (CJ) aims to improve the quality of performance-based assessments by letting multiple assessors judge pairs of performances. CJ is generally associated with high levels of reliability, but there is also a large variation in reliability between assessments. This study investigates which assessment characteristics influence the level of reliability. A meta-analysis was performed on the results of 49 CJ assessments. Results show that there was an effect of the number of comparisons on the level of reliability. In addition, the probability of reaching an asymptote in the reliability, i.e., the point where large effort is needed to only slightly increase the reliability, was larger for experts and peers than for novices. For reliability levels of .70 between 10 and 14 comparisons per performance are needed. This rises to 26 to 37 comparisons for a reliability of .90.

ARTICLE HISTORY
Received 12 April 2018
Accepted 26 March 2019

KEYWORDS
Comparative Judgement (CJ); Scale Separation Reliability (SSR); performance-based assessment; task-complexity; meta-analysis

Performance-based assessments, in which people show their competences in an authentic and largely uncontrolled environment, are generally considered to be a valid way of assessing competences (Lane & Stone, 2006). However, this kind of open-ended assessment also increases the complexity of the performance quality evaluation as there is no objective indication of what is better or worse. Instead, assessors have to make a subjective decision regarding the quality of performances, for which they commonly use their own experiences and their own standards as a basis to compare a particular performance with (Laming, 2003).

The method of comparative judgement (CJ) has the potential to improve the quality of evaluations in performance-based assessments by comparing the quality of performances directly with each other instead of evaluating them one by one in an absolute manner (Pollitt, 2004). CJ is considered to be easier and more intuitive, as people generally base their decisions on comparisons, either consciously or unconsciously (Laming, 2003). In an overview of studies using this comparative approach, presented in a research report, Bramley (2015) showed that CJ is generally associated with high levels of reliability. Yet, this overview also indicates a large variability between assessments, both in the way they are implemented (e.g., type of assessed performance or number of comparisons) and in the reliability of the results. Therefore, the current

CONTACT San Verhavert san.verhavert@uantwerpen.be Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium, Gratiekapelstraat 10, 2000 Antwerpen, Belgium

In the meantime Renske Bouwer changed affiliation. She is now affiliated with the Department of Pedagogical and Educational Sciences, Vrije Universiteit Amsterdam, the Netherlands

study aims to investigate which assessment characteristics influence the reliability of CJ assessment results. The goal of this is to provide research and practice with useful guidelines to set up CJ assessments.

The variation in reliability raises the question: when is an assessment reliable enough? The answer depends on the context and the type of the assessment. In the literature on the assessment practice, two boundaries are put forward, dependent on the main goal of the assessment. For low-stakes or formative assessments, in which learning is the main goal, reliability levels of .70 or higher are deemed sufficient (Jonsson & Svingby, 2007; Nunnally, 1978). For high-stakes and summative assessments in which important decisions are made on the basis of assessment results, the current study opted for the widely accepted reliability level of at least .90 (Nunnally, 1978). Second, in the context of research in (educational) assessment, the main goal of researchers is generally to reach a maximal level of reliability. Here, this maximal level of reliability does not refer to the value of 1. Rather, it is conceptualized as the asymptote. This is the point in the assessment where adding a few more comparisons will not cause in a significant increase in reliability. In other words, it is the point where large efforts (in number of comparisons) are required for only little gains in reliability. This leads to the additional question: is there an effect of assessment characteristics on whether a maximum level of reliability is reached (i.e., an asymptote)?

Before describing the methodology and results of the current study, the basic principles of CJ will be outlined, followed by a discussion of the key characteristics upon which assessments can be distinguished from each other.

Basic principles of comparative judgement

The idea of CJ originates from the field of psychometrics and is based on Thurstone's Law of Comparative Judgement (1927). It was introduced into the field of educational assessment as an alternative to conventional marking by Pollitt (e.g., 2004, 2012). The basic idea is that a group of assessors is asked to each individually judge which one out of two student works, further called representations, is better regarding the competence under assessment. By applying the Bradley-Terry-Luce (BTL; Bradley & Terry, 1952; Luce, 1959) model on a series of comparative judgements, it is possible to estimate the scaled values representing the quality of the representation. The scaled estimates are measured on a logit scale, reflecting the log of the odds that a particular representation is better than the 'average' one. It should be remarked that two assumptions are made in the BTL model (Bradley & Terry, 1952): (1) all representations can be distinguished from one another on a unidimensional continuum (here the quality of performance) and (2) all judgements and pairs are independent of each other.

In the past three decades, CJ has been applied to the assessment of a wide range of competences across the full educational spectrum, including primary (e.g., Heldsinger & Humphry, 2010), secondary (e.g. McMahon & Jones, 2015) and higher education (e.g. Steedle & Ferrara, 2016). In these assessments different types of performances are compared, such as written texts, audio, video and portfolios (e.g., Kimbell, 2007; Pollitt & Murray, 1995; van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016). Overall, this research shows reliabilities ranging from .73 (Jones & Alcock, 2014) to .98 (Heldsinger & Humphry, 2010). As there is quite some variance in the level of reliability

across assessments, the current study looks into what makes one set of results more reliable than others. Furthermore, recent research (Bramley & Vitello, 2018) shows that using adaptive algorithms to allocate pairs for judgment can spuriously inflate the reliability, which casts doubt on some of the reliability values that have been reported in adaptive CJ studies. It is therefore of great importance to further our understanding of the factors that affect the reliability of CJ.

The reliability of CJ: scale separation reliability

The reliability of CJ is most commonly determined by the Scale Separation Reliability (SSR) and is formulized as follows (Bramley, 2015; Wright & Masters, 1982):

$$SSR = \frac{G^2}{(1 + G^2)} \tag{1}$$

With

$$G = \frac{\sigma_\beta}{RMSE} \tag{2}$$

With σ_β the standard deviation of the estimates and *RMSE* the root mean squared error of estimation.

The SSR provides an indication of the internal consistency of the results of an assessment. Verhavert, De Maeyer, Donche, and Coertjens (2018) showed that the SSR is an estimation of the inter-rater reliability as well as of the split-half reliability. The SSR, thus, provides us with information on the reproducibility of the results across groups of assessors who are equivalent in background characteristics. It also informs on the consistency of the results across random halves of the assessor group. It can thus be stated that the SSR reflects the consistency between assessors in general. This could, however, also mean that judgements based on assessors who highly agree on what constitutes a good performance and on what are quality criteria for such performances, are associated with a higher SSR value and possibly also a lower validity (Lesterhuis, 2018; see also Lesterhuis et al., 2018; van Daal et al., 2016).

The SSR can thus be seen as a measure for assessor accuracy or consistency. Inter-rater reliability can therefore be defined as consensus in judgement (Stemler, 2004). It is also possible to measure consensus or consistency by letting all assessors judge a common set of pairs and then calculating the agreement on these pairs. One disadvantage is that this agreement should be calculated for each pair of judges (Stemler, 2004). The relative usefulness of this and other forms of measuring inter-rater consensus is a topic of further investigation.

Like reliability in Classical Test Theory (CTT) the SSR can be conceptualised as a ratio of true variance to (true + error) variance. In CTT this leads to the observation that the reliability of a test can be increased simply by administering it to a group of more widely varying ability levels. This is also true with the BTL model (Bi, 2003). Therefore, it can be argued that, like in CTT, it might be better to use the SEM to evaluate assessment quality (e.g., Tighe, McManus, Dewhurst, Chis, & Mucklow, 2014).

In the current article we were interested in assessor accuracy and consistency, so we opted to use the SSR as quality criterion. Specifically, in the context of CJ, if representations are closer together in ability score (lower true variance), it will become more

difficult for assessors to accurately discriminate between the quality of the representations. This then leads to a lower SSR (Bi, 2003). However, if the assessors are better at discriminating representations, a low true variance can still result in a good SSR value. Given that the effects of assessor discrimination and true variance of representations are usually confounded within an individual study, it makes sense to focus on the SSR.

Minimum number of comparisons

The most prominent characteristic of a CJ assessment is the number of comparisons. CJ assessments require a large number of comparisons for a high level of reliability. As with most statistical models, more information will make the estimates in the BTL model more accurate and reliable. In addition, the more comparisons assessors make, the more familiar they will become with the comparative task and the representations. According to the model of task complexity of Liu and Li (2012), decreased novelty of the input, i.e., the representations and the task, will decrease the overall task complexity, resulting in more consistent and hence, reliable judgements. On the other hand, Bramley, Bell, and Pollitt (1998) reported that judges often find the task tedious and time consuming, which can negatively impact task performance, i.e., through fatigue (Liu & Li, 2012). It might thus be beneficial to stop at a minimum number of comparisons when reaching a sufficient level of reliability.

Bramley (2007) and Pollitt (2012) argued that in CJ a minimum threshold exists at which each object receives enough judgements to reliably calculate logit scores. However, they have not explicated the exact level of this threshold. So far, only one study reported 10 comparisons per representation as an empirical minimum in random CJ (Wheadon, 2015). As this value was based on a post-hoc simulation study in which resampled results from parts of the full data set were correlated with the eventual scale, the actual level of reliability was not taken into account. Hence, no information is yet provided on the measurement accuracy for CJ as expressed by the reliability measure. The current study attempts to bridge this gap by investigating how many comparisons are needed for specific levels of reliability in random CJ.

Other assessment characteristics

A second prominent characteristic of CJ is the number of assessors. One of the strengths of CJ is that it allows for, and even promotes, the use of multiple assessors, which increases the validity and generalizability of the assessment results (Lesterhuis et al., 2018; van Daal et al., 2017). It could also be argued, however, that engaging more assessors in the assessment leads to more differing perspectives on the particular competence under assessment, which might decrease the reliability. Up until now it is unknown to what degree results are affected by judges with a deviating opinion regarding the competence under assessment. Furthermore, when the number of comparisons is kept constant, increasing the number of assessors will decrease the number of comparisons per assessor. As stated in the previous section, this might negatively influence the reliability of the results through the reduced familiarity of the assessor with the representations and the judgement task. Specifically, when assessors have to complete only a small number of comparisons, the representations and the CJ task will

remain relatively novel and complex which makes it harder for assessors to provide consistent judgements (cf. Liu & Li, 2012). In contrast, more comparisons per assessor will likely lead to more consistent, and hence reliable, judgements due to decreased complexity of the overall judgement task.

A third assessment characteristic that may affect the level of reliability in CJ is the format of the assessed product. Some competences can be assessed with different kinds of products, including written texts, audio, video, and portfolios (Kaslow et al., 2007; McMullan et al., 2003). These formats differ from each other in the structure, ambiguity and diversity of the presented information. These differences might have an influence on the task complexity (Liu & Li, 2012) and, hence, on the reliability of the overall results.

A fourth characteristic is whether and how assessors provide feedback to students. Feedback is considered to be very important to improve learning (Dochy & McDowell, 1997). But how can feedback be implemented in CJ assessments? One possibility is to ask for feedback after the CJ assessment is completed. However, this is not a very efficient method, as assessors have to evaluate (part of) the products once again. Another possibility that is potentially more efficient is to integrate the feedback into the judgement flow by asking for feedback after each comparison (Van Gasse et al., 2017). However, in traditional marking it appears that it is implicitly assumed that feedback requires analytic judgements, instead of holistic judgements(e.g., Bacha, 2001; Foltz, Gilliam, & Kendall, 2000; Sadler, 2009). Thus, providing feedback and making holistic comparative judgements might be two completely different processes, requiring different mindsets. As a consequence, implementing feedback during the flow can negatively affect the complexity of the judgement task, and hence, the reliability of the results (Liu & Li, 2012).

A fifth characteristic is how many representations that have to be assessed. To our knowledge, it is not yet known what the effect is of the number of representations on the outcomes of a CJ assessment. A tentative hypothesis is that more representations comprise a higher range in quality, and hence a higher true variance. Due to the comparative nature of CJ, larger quality differences may facilitate the differentiation between representations (van Daal et al., 2017), leading to more consistent judgements and, hence, a higher reliability. Furthermore, more representations imply that assessors will have to make more comparisons, if the number of assessors is fixed. As previously stated, more comparisons might decrease the novelty of the task and the input, leading to a less complex task (Liu & Li, 2012) and hence, to a higher reliability.

The sixth and final assessment characteristic is the expertise of assessors. Comparisons can be made by expert judgements, but also by peers or novices. Previous research has indicated that CJ can be used as a valid and reliable peer assessment tool (Goossens, Bouwer, & De Maeyer, 2017; Jones & Alcock, 2014). In two studies by Jones and colleagues (Jones & Alcock, 2014; Jones & Wheadon, 2015), validity was defined as the correlation between the results of peers or novices on one hand and the results of experts on the other hand. These studies showed that peers can reach results that are as reliable and as valid as those of experts in CJ. Novices reached results with a lower validity but equal reliability compared to experts and peers. It is, however, possible that experts, peers and novices reach this reliability level at different speeds, expressed in numbers of comparisons. It could be hypothesized that, compared to novices, experts (and peers) have a better understanding of the competence that is

assessed and the relevant elements for a quality performance. Therefore, experts (and peers), compared to novices, might reach consensus, i.e., high levels of reliability, after less comparisons.

There are still other, but less prominent, characteristics that may affect levels of reliability in CJ. For example, in some assessments assessors have to make judgements on more than one competence or criteria (e.g. Humphry & Mcgrane, 2015). Furthermore, the reliability may also be affected by the (a priori) exclusion of incomplete or inconsistent representations or the post hoc exclusion of deviating judges. These characteristics were not included in this study because they did not appear in the data set and could less be considered as main defining characteristics of (CJ) assessments.

Current study

The current study aims to provide insight in the systematic causes of variability in the reliability of CJ assessment results as expressed by the SSR. More specifically, it looks into the effects of the following assessment characteristics: the number of comparisons, the number of assessors, the format of the representations, feedback, the number of representations, and assessor expertise. This leads to the first research question:

RQ1. What is the effect of assessment characteristics on the reliability of the assessment?

The current study not only aims to explain the variability in the reliability but also to provide practitioners with guidelines for setting up CJ assessments in such a way that reliable results are obtained in the most efficient way. The desired level of reliability largely depends on the purpose of the assessment. In the context of research and for certain assessments, it is important to maximize the level of reliability. In this study this maximum level in the reliability is conceptualized as the asymptote in the reliability. Specifically, this is the point in the assessment where large efforts will be needed in order to further increase the reliability. Therefore, the second research question is:

RQ2. Which assessment characteristics increase the probability of reaching a maximum level of reliability (i.e., asymptote of SSR)?

In order to answer these questions a meta-analysis was conducted. In the methods section the authors first provide a description of the data regarding the major assessment characteristics and some derived measures. The methods section concludes with a description of the analysis procedure.

Method

The data

This study used data from 49 CJ assessments conducted between 2014 and 2016. All assessments used the Digital Platform for the Assessment of Competences (D-PAC; www.d-pac.be) an online CJ assessment tool. In D-PAC, pairs are automatically generated by a distributed random algorithm, in which each representation is compared multiple times to another object. This algorithm constructs pairs of single representations based on two criteria: (1) representations have been compared the smallest number of times and (2)

representations have not yet been paired with each other. If more than one representation meets these criteria, representations are randomly chosen. For clarity, this algorithm is based on randomness and is thus not adaptive.

The CJ assessments were conducted in different contexts: primary (n = 7), secondary (n = 10) and higher (n = 18) education, research (n = 7), and selection (n = 7); with the majority in higher education. Assessments included a wide variety of competences in different representation formats. The average reliability over all assessments was .78, ranging from .45 to .99. Tables 1 and 2 summarize the main assessment characteristics of the assessments. For a detailed overview of the assessment characteristics see the supplementary materials with this article. The full dataset is also made available by the authors through the Zenodo repository (Verhavert, Bouwer, Donche, & De Maeyer, 2018).

Number of comparisons (N_C) and number of representations (N_R)

On average, an assessment involved 817 comparisons (min. = 54, max. = 9038) and 84 representations (min. = 6 max. = 1089). In CJ the number of comparisons is commonly presented as *the number of comparisons per representation* (N_{CR}; M = 28, min. = 7, max. = 297). This is calculated using the following formula: $(N_C/N_R) \times 2$. We multiply by two because one comparison consists of 2 representations.

The proportion of the full matrix is another way to represent the total number of comparisons. A full matrix includes all possible pairs for this group of representations and its size is calculated by $[N_R \times (N_R - 1)]/2$. It thus represents the maximal amount of information that can be obtained.[1] The proportion of the full matrix thus expresses how far along the assessment is towards maximal information. It was, however, decided not to include this measure because it is closely related to the number of comparisons per representation. This is also illustrated by the high correlation between the two (r = .88; Table A1 in Appendix A).

Table 1. Mean and Standard Deviations of Number of representations (N_R), number of assessors (N_A) and number of comparisons (N_C) per assessment context.

	N_R		N_A		N_C	
Context	*M*	*SD*	*M*	*SD*	*M*	*SD*
Primary Education	164	74	32	42	1283	611
Secondary Education	79	47	25	16	785	608
Higher Education	39	30	36	37	413	260
Research	187	399	30	32	2015	3251
Selection	27	15	13	11	239	133

Table 2. Number of assessments per assessment context by representation media type, level of assessor expertise, and feedback type.

	Representation Media Type					Level of Assessor Expertise			Feedback Type		
Assessment Context	Text	Image	Audio	Video	Portfolio	Experts	Peers	Novices	None	Comparative	Pro-Con
Primary Education	7	0	0	0	0	1	0	6	4	0	3
Secondary Education	9	1	0	0	0	9	0	1	2	5	3
Higher Education	11	3	0	4	0	6	8	4	3	1	14
Research	0	1	3	3	0	6	0	1	3	1	3
Selection	5	0	0	0	2	4	1	2	0	3	4

Due to the general set-up of the assessments in D-PAC a large correlation can be observed between N_C and N_R (r = .94; Table A1 in Appendix A). That is, the *number of comparisons* (N_C) was calculated by multiplying N_R with a fixed number, i.e., N_{CR}. This correlation might cause collinearity or redundancy effects in the analyses. However, as N_C and N_R are two essential characteristics of CJ assessments, both are included in the analyses.

Number of assessors (N_A)

The *number of assessors* is a corrected value. This means that assessors were only counted if they completed at least 1/3 of the planned number of comparisons each assessor had to make. Assessments included 29 assessors on average (min. = 4, max = 127). Based on N_A and N_R the data can be split into two distinct groups: (1) regular assessments with a few assessors for a lot of representations and (2) peer assessments with a comparable number of assessors and representations. Therefore, also the *number of representations per assessor* (N_{RA}; M = 5, min. = 1, max. = 21) was calculated.

Format of representations

In most assessments (n = 32) the representations were *texts*, either typed or hand-written. The other assessments consisted of *images* (n = 5), *audio* or *video* (n = 10), or *portfolios* consisting of multiple media types (n = 2). For the analyses representation format was recoded into dummy variables with *text* as the reference category.

Feedback

In D-PAC *feedback* is included in the judgement flow. In other words, assessors were asked for feedback after every comparison. To find an optimal balance in the amount of feedback students receive (Cho & Schunn, 2018), some assessments required feedback for only a part of the comparisons. Assessments were coded for the inclusion of feedback if feedback was asked for more than half of the comparisons. This resulted in 37 assessments. There were two types of feedback: *comparative feedback* and *pro-con feedback*. For *comparative feedback* (n = 10), assessors were asked to briefly explain their choice. For *pro-con feedback* (n = 27), assessors had to indicate positive aspects as well as aspects that needed improvement for each representation separately. *Feedback* was also dummy coded for the analyses, with *no feedback* as the reference category.

Assessor expertise

Assessor groups consisted of *expert assessors* (26 assessments), *peers* (9 assessments) or *novices* (14 assessments). *Expert* judges were defined as judges who were experts (in assessment) in the field, and/or who received specific training in the particular (CJ) assessment task. *Peer* assessments were only conducted in the context of higher education and selection. In higher education, *peers* were fellow students who had performed the task under assessment themselves or had experience with the task in a previous phase of their studies. In the context of selection, *peers* were fellow researchers. *Novices* had no domain-specific expertise, and/or no experience in assessment in that field, and they did not receive any training beyond general instructions on how to use the digital

tool. The variable *expertise* was dummy coded for the analyses with *experts* as the reference category.

Scale separation reliability (SSR)

The reliability is calculated with the formula for *SSR* as presented earlier in this paper. In the data of the current study the mean *SSR* is .79 (min. = .50, max. = .99) and generally clusters between and around .70 and .80.

The *SSR* value is also calculated per round, in order to analyse the effects of assessment characteristics on the probability that an asymptote is reached. A round specifies the times that each representation is compared, i.e., the number of the round is the same as the number of comparisons per representation. For the probability that an assessment reaches an asymptote (RQ2) the difference in *SSR* between the current and the previous round (returns on investment) is calculated (grey points and line in Figure 1). An assessment is coded as having reached an asymptote if the returns are below the critical value of .01 for three consecutive rounds.

Data analysis

The current study attempts to identify which assessment characteristics influence the level of reliability (RQ1) and the probability of reaching an asymptote in reliability (RQ2). For clarity and transparency of this exposition, we discuss the main results based on descriptive figures and tables. The results are, however, confirmed statistically by performing regression analyses based on a two-phase procedure. The first phase consists of a classical or frequentist approach in which a forward stepwise regression procedure is used to identify possible effects using the Bayesian Information Criterion (BIC). In essence, this procedure identifies the best fitting

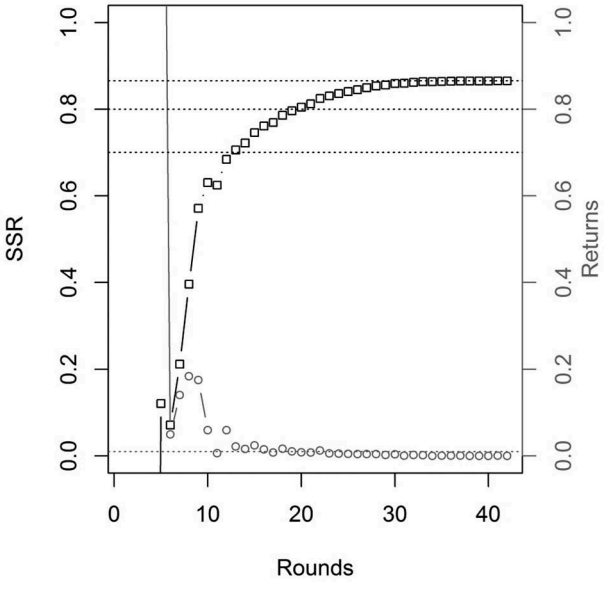

Figure 1. The reliability (*SSR*; black line and squares) and returns on investment (*Returns*; grey line and circles) per *Round* (i.e. number of comparisons per representation).

regression model by selecting those variables that produce the largest drop in BIC (ΔBIC). In the second phase, the parameters in the best model from phase one are estimated using Bayesian statistics. This approach was chosen because Bayesian statistics express how credible values for regression parameters are, which is more intuitive to interpret. For ease of understanding and in order to make some suggestions for practice, posterior predictions were made, based on the Bayesian modelling. All analyses are done in R (R Core Team, 2017).

Before moving to the results, the regression models will shortly be discussed. For details regarding the analysis and the Bayesian models, interested readers are referred to the supplemental materials with this article.

In RQ1 the dependent variable is the level of reliability. This value is, like every reliability index, theoretically limited to the interval $]0;1[$. It follows a normal distribution bounded on $]0;1[$ (Figure 2.a.) and appears to have a sigmoidal relationship with characteristics such as the number of comparisons per representation (Figure 2.b. and 2.c.). Therefore, to analyse which assessment characteristics have an effect on the level of reliability, the regression consisted of a generalized linear model with a Gaussian family and a logit link function.

In RQ2, the dependent variable is binary, the reliability reaches an asymptote (= 1) or not (= 0). To estimate the probability of reaching an asymptote, depending on assessment characteristics, a classical logistic regression is used.

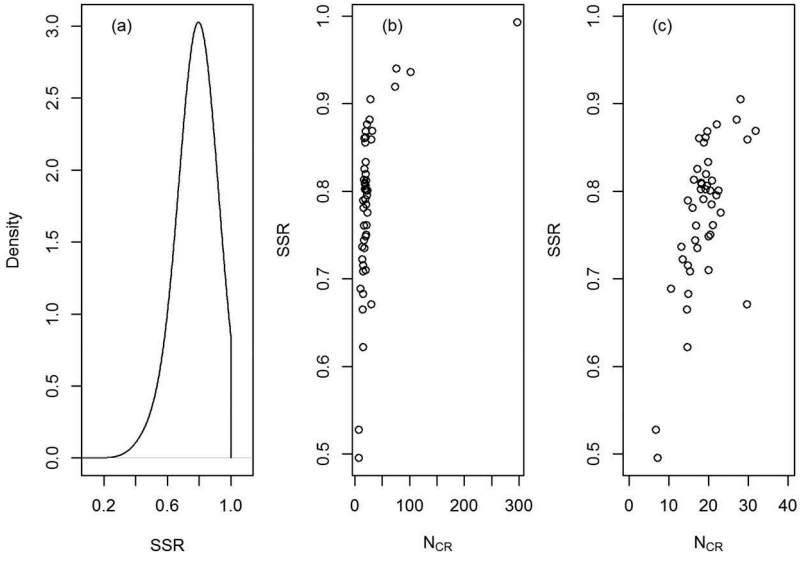

Figure 2. *(a)* Density plot of the reliability (*SSR*), *(b)* a plot of reliability (*SSR*) by number comparisons per representation (N_{CR}) and (c) a detail of the reliability (*SSR*) by number of comparisons per representation (N_{CR}).

Results

RQ1: effect of assessment characteristics on the reliability

Figure 3 and Table 4 show the relationship between the different assessment characteristics and the *SSR*. From Figure 3 it can be observed that only the *number of*

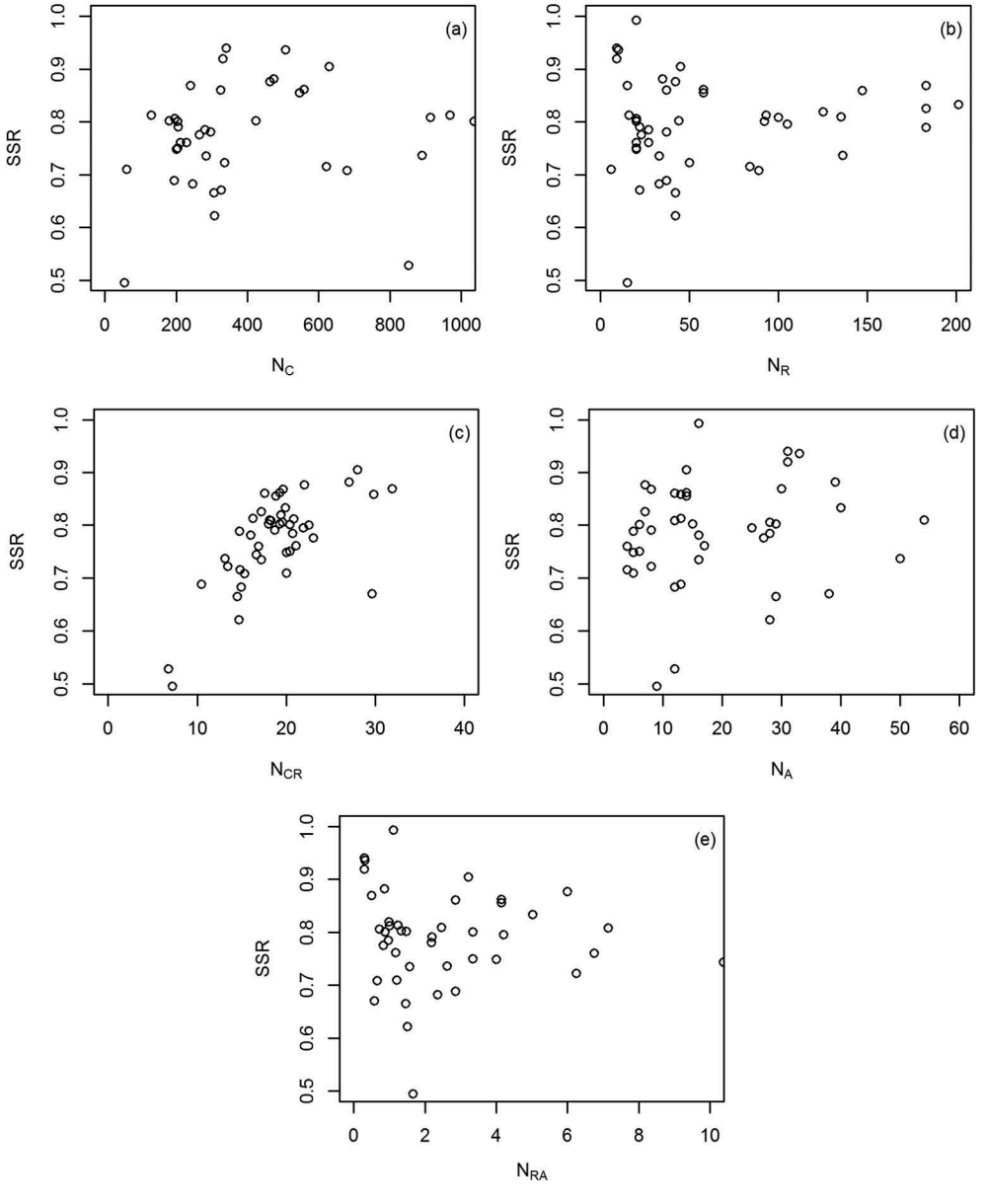

Figure 3. Relationship between reliability (*SSR*) and (a) number of comparisons (N_C), (b) number of representations (N_R), (c) number of comparisons per representation (N_{CR}), (d) number of assessors (N_A) and (e) number of representations per assessor (N_{RA}). For clarity each graph only shows part of the data.

Table 4. Descriptives per representation type, level of assessor expertise and feedback type.

	n	N_{CR} *M (SD)*	N_{RA} *M (SD)*	*SSR* *M (SD)*	Percentage asymptote
Representation Type					
Text	32	18 (4)	6 (9)	.77 (.08)	28.13%
Image	5	77 (123)	4 (4)	.83 (.10)	40.00%
Audio	3	21 (6)	7 (4)	.81 (.08)	66.67%
Video	7	47 (37)	1 (2)	.80 (.16)	85.71%
Portfolio	2	20 (NA)	3 (NA)	.77 (.04)	0%
Assessor Expertise					
Experts	26	19 (6)	5 (6)	.79 (.08)	46.15%
Peers	9	19 (2)	1 (0)	.75 (.05)	22.22%
Novices	14	51 (76)	7 (11)	.80 (.13)	35.71%
Feedback Type					
None	12	48 (82)	7 (7)	.80 (.12)	50.00%
Comparative	10	25 (18)	3 (2)	.82 (.11)	50.00%
Pro-Con	27	21 (12)	5 (9)	.77 (.08)	29.63%

Note. *n* = the number of assessments in this level, *M* = mean, *SD* = standard deviation, N_{CR} = number of comparisons per representation, N_{RA} = number of representations per assessor, *SSR* = scale separation reliability, Percentage asymptote = percentage of assessments in this level that reach asymptote.

comparisons per representation (N_{CR}; Figure 3.c) has an effect on the *SSR*. Other characteristics such as the *number of comparisons*, the *number of representations* or the *number of assessors* do not have an effect on the SSR. This was confirmed by the forward stepwise regression procedure. Namely, adding the *number of comparisons per representation* to the null model led to a reduction in BIC of 43.49. No additional variables caused a further reduction larger than 2.

With the interpretation of Table 4, cautiousness is warranted regarding *representation type*. Namely, because the number of assessments in each category is small compared to the category *text*, it is unwarranted to draw any conclusions from the raw data. For the categories of *assessor expertise* and *feedback type* this is not a problem. For these two variables it can be observed that the means (*M*) of the *SSR* do not differ between levels and types when the standard deviation (*SD*) is taken into account. However, it is possible that this is an artefact due to a difference in mean N_{CR} between the categories. The overlap in N_{CR} between the categories, as shown by the *SD*, is large enough to say that N_{CR} is probably not a confounding factor. Again, the forward stepwise regression procedure confirmed that there was no effect of *assessor expertise* and *feedback type*. It can thus be concluded that only N_{CR} influences the reliability level.

Based on the Bayesian parameter estimates, presented in Table 5, a posterior prediction was performed (Figure 4) to see what values can be expected for specific levels of the *SSR*. The results of the posterior prediction show that an average of 13 comparisons per representation (min = 10, max = 14) is associated with a reliability of .70. A reliability of .80 is reached between 19 and 20 comparisons per representation. For a reliability of .90, i.e., with high

Table 5. Bayesian estimates and credible intervals for multiple regression analysis for the effect of assessment characteristics on the SSR.

Predictor	Posterior median	95%-HDI
Intercept	2.06	[1.73; 2.39]
Comparisons per Representation (N_{CR})	.08	[.05; .11]

NOTE: 95%-HDI = 95% highest density interval.
All values in this table are expressed in logit values.

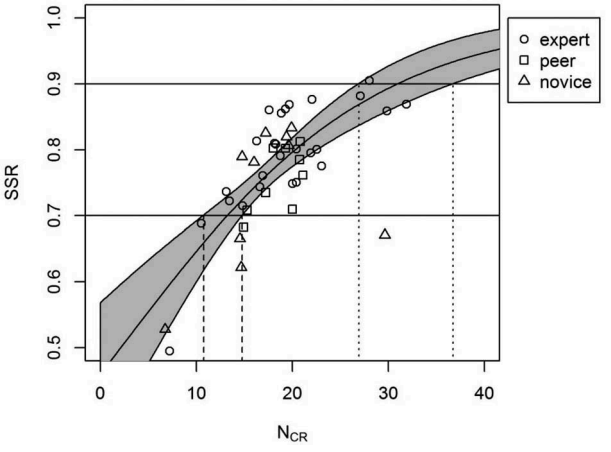

Figure 4. Posterior prediction. Reliability (*SSR*) predicted by the number of comparisons per representations (*N[CR]*; black line) with 2 SD uncertainty interval (Shaded area) plotted beside the data divided by assessor expertise.

stakes assessments, between 26 and 37 comparisons are needed. The ranges of these predictions differ quite a lot. This is because in Bayesian analysis, more data means less uncertainty in the estimates, thus a higher accuracy. As can be observed from Figure 4, the data (circles, triangles and squares in the plot) are mainly clustered around *SSR*'s of .70 and .80. It can also be observed that there is an outlier in the data around 30 comparisons per representation. Some further checks have shown that this outlier negligibly influences the results.

RQ2: effect of assessment characteristics on the probability of asymptote

To visualize which characteristics influence the *probability that an asymptote is reached* in the reliability, the following steps were performed. The assessment characteristics that were continuous variables were divided into equally spaced intervals and the percentage of assessments reaching an asymptote was calculated and plotted per interval (Figure 5). Again, a clear effect is observed for the *number of comparisons per representation*. Additionally, the *number of representations per assessor* also shows a relationship with the *probability to reach an asymptote*. For the other characteristics, *number of comparisons, number of representations* and *number of assessors*, there are no effects. The forward stepwise regression procedure and Bayesian estimates confirmed these results (Tables 6 and 7).

For the categorical assessment characteristics the calculated percentages are presented in Table 4. The same remark as with RQ1 applies here. Namely that the categories for *representation type* contain too few assessments to make any solid conclusions. Regarding *assessor expertise* there appears to be a difference between the three categories. *Peers* appear to have the lowest percentage of assessments reaching an asymptote after *novices*. The difference between *peers* and *expert* could however be due to the difference in *number of representations per assessor*. Although there is an overlap in *number of representations per assessor* from the side of *experts*. This difference in percentage should thus be confirmed by the analyses.

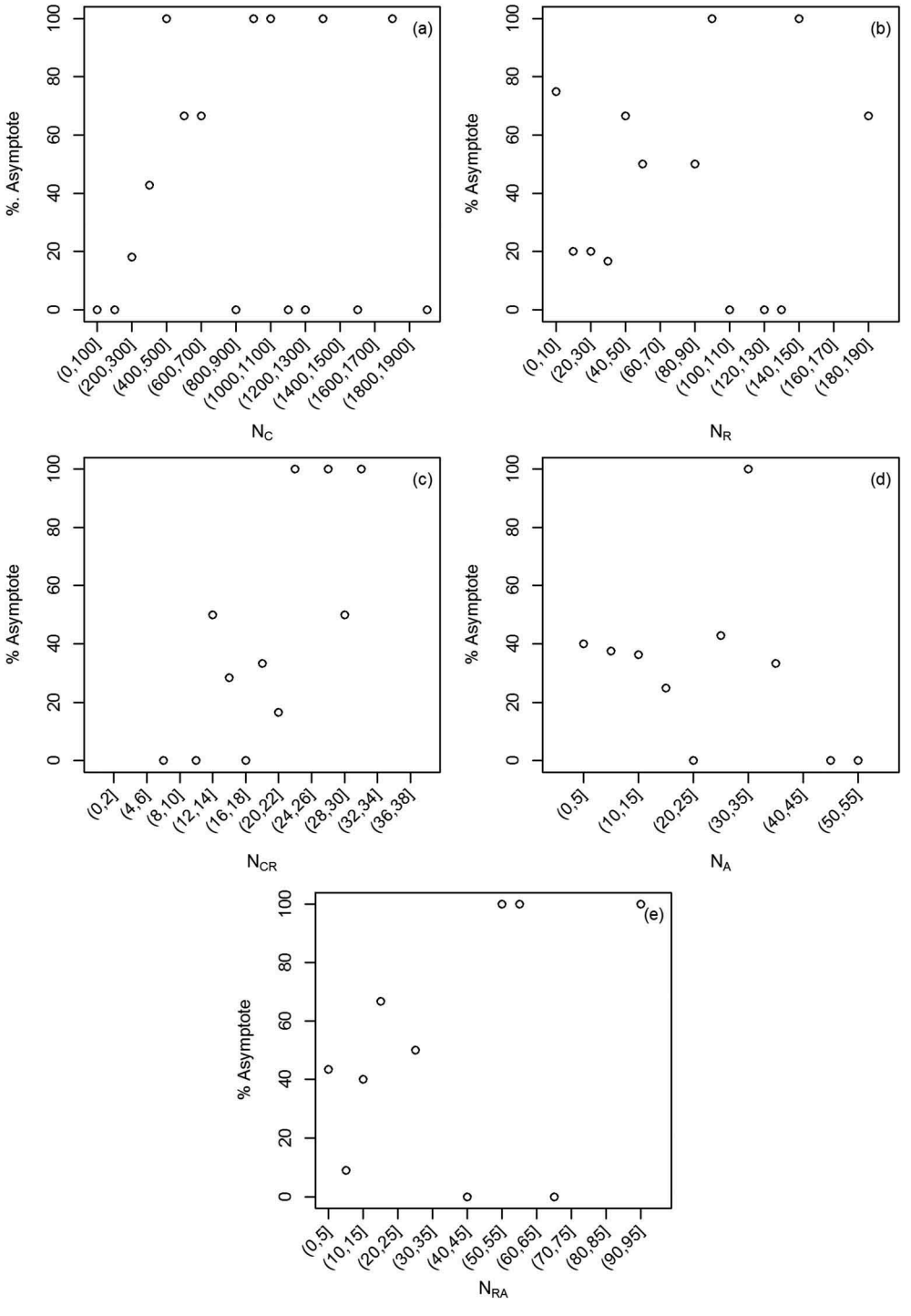

Figure 5. Relationship between the percentage of assessments reaching an asymptote (*% Asymptote*) in fixed width intervals of (a) number of comparisons (N_C), (b) number of representations (N_R), (c) number of comparisons per representation (N_{CR}), (d) number of assessors (N_A) and (e) number of representations per assessor (N_{RA}). For clarity each graph only shows part of the data.

Table 6. BIC and drop in BIC for multiple regression analysis predicting the chance of reaching an asymptote in the SSR from assessment characteristics.

Predictor	BIC	ΔBIC [a]
Intercept	69.33	
Step 1		
Comparisons per Representation (N_{CR})	57.55	11.78
Step 2		
Representations per Assessor (N_{RA})	54.28	3.27
Step 3		
Expertise	51.17	3.11
n	49	

NOTE: [a] ΔBIC = drop in BIC. The difference between the current BIC and the BIC in the previous step.

Table 7. Bayesian estimates and credible intervals for multiple regression analysis predicting the chance of reaching an asymptote in the SSR from assessment characteristics.

Predictor	Posterior median	95%-HDI
Intercept	3.14	[−0.52; 6.24]
Comparisons per Representation (N_{CR})	0.39	[0.12; 0.74]
Representations per Assessor (N_{RA})	0.30	[0.08; 0.60]
Expertise – peers	0.20	[−3.41; 1.85]
Expertise – novices	−5.52	[−27.45; −1.78]

NOTE: 95%-HDI = 95% highest density interval.
All values in this table are expressed in log odds.

The forward stepwise procedure confirmed the effect of *expertise* (Table 6). Furthermore, the Bayesian estimates show that only *novices* differ from *experts* and *peers* do not (Table 7). It is thus probable that the difference in Table 4 is indeed due to the *number of representations per assessor*.

From the posterior prediction (Figure 6) it can be observed that with five representations per assessor (the mean in the data set; Figure 6.a), it is very unlikely to reach an asymptote with 10 to 14 comparisons per representation. In other words, when most assessments reach a reliability of .70, they have not yet reached their asymptote. For the number of comparisons per representation to reach a reliability of .90, between 26 and 37, it can be observed that almost all assessments are predicted to have reached an asymptote. This is however only true for assessments with *expert* and *peer* assessors, as there are no differences between *experts* and *peers*. However, assessments with *novices* need more *comparisons per representation* and *representations per assessor* for an equal probability on reaching an asymptote. Therefore, with 37 comparisons per representation, the predicted *probability of reaching an asymptote* in assessments with *novices* is around 80%.

The influence *of number of representations per assessor* is most interesting in the planning phase of assessments. Specifically, if an assessment comprises a low number of representations per assessor, as is the case for peer assessments, it appears to be better to plan a larger number of comparisons per representation if the goal is to get the maximum out of an assessment (Figure 6.b).

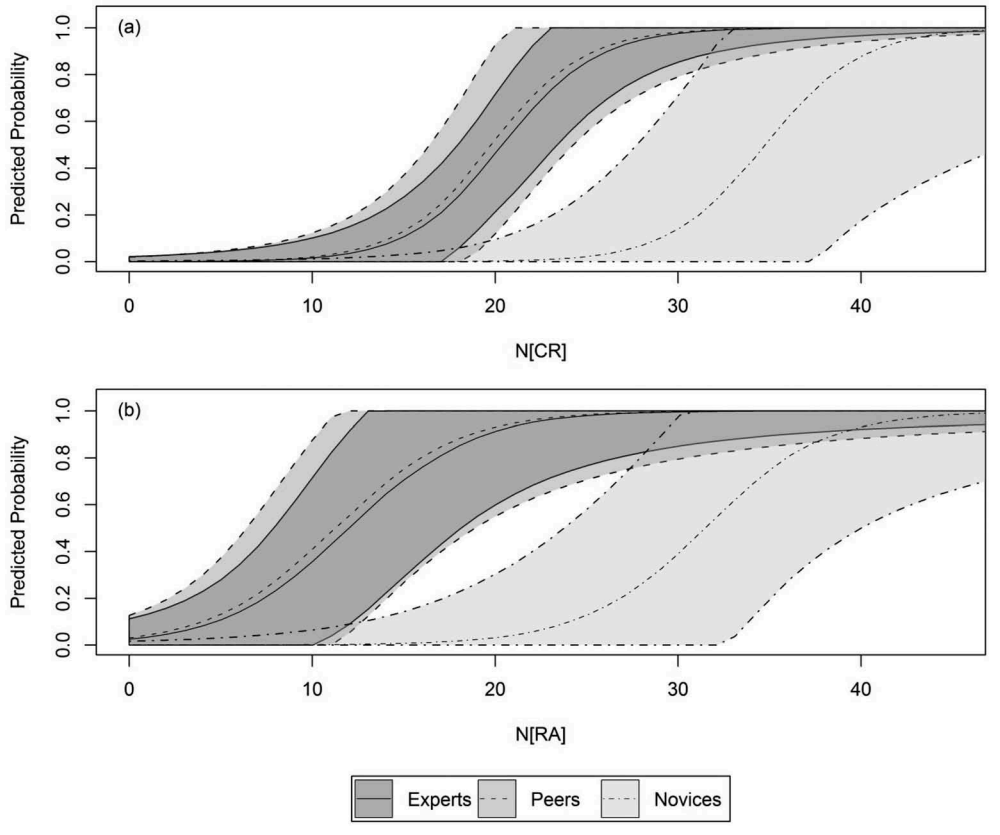

Figure 6. Posterior prediction. *Predicted probability* to reach an asymptote in the reliability by number of comparisons per representation (*N[CR]; a*) and by number of representations per assessor (*N[RA]; b*) for *Experts* (solid line), *Peers* (dashed line) and *Novices* (dot-dashed line). The shaded areas indicate the 2 *SD* uncertainty intervals.

Discussion

To better understand the mechanisms behind CJ and to account for an accurate and efficient implementation of CJ in both research and practice, the current meta-analysis investigated the effects of assessment characteristics on the reliability of a CJ assessment. This meta-analysis was based on 49 CJ assessments that were highly different in terms of the context in which the assessment took place, but also in assessment characteristics such as the number of comparisons in total and per representation, number of judged representations, the number and expertise of assessors, the number of representations per assessor, and the format of the representations and the feedback, as well as the reliability of the assessment results. There were two research questions central in this study. In the first RQ, the effects of assessment characteristics on the average level of reliability were examined. As researchers and practitioners could also be interested in the maximum level of reliability for credible results and how efficiently those can be reached, we examined the effects of assessment characteristics on the probability that a maximal level of reliability is reached (RQ2).

The overall results showed that the number of comparisons per representation is the only characteristic that consistently affects the reliability across both analyses. The analyses predicted that between 10 and 14 comparisons per representation are needed to reach a reliability of .70. To reach a reliability of .90, 26 to 37 comparisons per representation are needed. Wheadon (2015) proposed 10 comparisons per representation, based on correlations and post-hoc simulations. This value lies within the range of the predicted reliability of .70. However, because Wheadon (2015) does not report any level of reliability it is difficult to fully compare these results. When the aim is to reach a level of reliability that is as high as possible, the current study predicted that a lot more comparisons per representation are necessary. Specifically, with five representations per assessor and 10 to 14 comparisons per representation, the expected probability to reach an asymptote is lower then 20%. For a probability of 90% to reach an asymptote, between 20 and 35 comparisons per representation are needed if assessors are experts or peers. If they are novices than between 30 and 45 comparisons per representation are needed.

Besides the confirmed effect of the number of comparisons per representation, it is striking that quite some expected effects were not confirmed. It appears that it is possible to reach reliable results with a CJ assessment regardless of the number of representations included in the assessment. Also, providing feedback during the assessment does not have an influence on the level of reliability, although it might affect the holistic character of the judgements. Specifically, it is implicitly assumed that feedback can only be provided with analytic assessment (e.g., Bacha, 2001; Foltz et al., 2000; Sadler, 2009). It could thus be assumed that feedback forces assessors to adopt an analytic strategy which might conflict with the (assumed) holistic character of CJ. Based on the findings of this study, future research could investigate if CJ really consists of holistic judgements. Furthermore, it can be studied if asking judges for feedback makes their assessments more analytic and if this influences the results.

It is especially interesting that either number, nor level of expertise of the assessors appeared to matter in how reliable an assessment can be. This implies that it is possible to increase the validity (in the sense of involving a wider range of opinions) by adding more assessors, without harming the reliability of the assessment. However, it can be argued that including more assessors might increase the chance of having deviating assessors. Moreover, in her dissertation Lesterhuis (2018) has shown that assessors can differ in their opinion on aspects upon which representations are compared. This might influence the results, but can also increase the validity, with the assessment results providing more comprehensive image of the competence. However, it is not currently known whether and how assessors who misfit the BTL model significantly affect the reliability. Further research on this topic is needed.

The results showed that assessor expertise influences how much effort assessors need to put into the assessment if the goal is to get as much out of the assessment as possible. It appeared that for novices a larger number of comparisons per representation are required for a maximum reliability than for experts and peers. This finding in part confirms the results of Jones and colleagues who found a significant difference in reliability between expert and novice assessors but not between peer and expert assessors (Jones & Alcock, 2014; Jones & Wheadon, 2015). An explanation for this finding can be that experts as well as peers are familiar with the representations and know what to look for in an assessment, increasing the

consistency between assessors. It also makes the process of comparative judgement of the representations less cognitively demanding (Liu & Li, 2012). Lower cognitive demands might lead to a higher consistency within assessors. With a higher consistency between and within assessors less information (i.e., comparisons) is needed to reach maximal levels of reliability. An important remark here is that expertise does not limit the level of reliability that can be achieved, i.e., novices can reach as high levels of reliability as experts and peers.

In sum, the results lead to the following practical implications for research and practice. When setting up a CJ assessment the main focus should be on the number of comparisons per representation. When the assessment is formative, i.e., results are used for learning purposes, between 10 and 14 comparisons per representations is needed in order to reach a reliability level of .70. When aiming for a reliability level of .90 in order to make summative decisions like a pass or fail, 26 to 37 comparisons per representation are needed. It is still recommended to regularly check the level of the reliability during the assessment and to add more comparisons when needed. It is generally recommended to use experts or peer assessors, especially if the aim is to get the maximum out of an assessment in the most efficient way.

The current study focussed on possible influences of CJ assessment characteristics on inter-rater consistency. Therefore, it was opted to use the SSR as measure of accuracy and consistency. As was remarked in the literature section, there is another commonly used measure of accuracy, namely SEM (Tighe et al., 2014). It was argued there that, in CJ, SEM can, theoretically, be seen as a consistency of accuracy in the estimates. This is certainly also an interesting and important measure to look into. Therefore, it can be recommended for further research to replicate the current study with SEM as dependent variable. Also in the context of inter-rater consistency, there exists some alternative measures, like assessor agreement on pairs that are the same for all assessors (Stemler, 2004). It might therefore be interesting to investigate the worth of these measures and in a later stage their potential variance.

It should be remarked that, although the results of this meta-analysis are based on an extensive amount of assessment, the findings are yet explorative. To draw firm conclusions on the effects of assessment characteristics on reliability in CJ, it is recommended to replicate and extend these results with a more experimental control over the variables of interest. It should be remarked that all assessments were conducted with the D-PAC platform and a random pair construction algorithm. Therefore, it might be possible that on different platforms and or with different pair construction algorithms results may differ. On the other hand, similar results may be expected when pair construction algorithms are similar to the one implemented in the included D-PAC assessments.

The data and the R-code used are made available through the Zenodo repository (Verhavert et al., 2018). Therefore, the results can be replicated by extending the data with assessments in different, international contexts. Furthermore, it is possible to test different and/or more complex models like mediation analysis. The availability of the data also makes it possible to test relationships between the assessment characteristics. On top of that, the more or less explicit assumptions in the Bayesian models can be checked.

Note

1. The maximum information that can be obtained if each pair is compared once. It might be possible to increase the information by judging each pair multiple times. This should however be checked.

Acknowledgments

The authors want to thank the two reviewers and the editor for their critical and helpful remarks. They helped increase the clarity and quality of this paper.

The assessments used in the data were conducted within and outside the University of Antwerp and with the cooperation of the following persons: Prof. dr. Wilfried Admiraal (Leiden Univerisity), Prof. dr. Kris Aerts (KULeuven), Prof. dr. Michael Becker-Mrotzek (University of Cologne), Nathalie Boonen (CLiPS University of Antwerp), Pia Claes (University of Cologne), Ilke De Clerck (CLiPS University of Antwerp), Liesje Coertjens (Université Catholique de Louvain), Cynthia De Bruycker (Hasselt University), Tinne De Kinder, Fien de Smedt (Ghent University), Prof. dr. Benedicte de Winter (University of Antwerp), Prof. dr. Steven Gillis (CLiPS University of Antwerp), Evghenia Goltsev (University of Cologne), Maarten Goossens (University of Antwerp), Ann-Kathrin Hennes (University of Cologne), Prof. dr. Hanne Kloots (CLiPS University of Antwerp), dr. Marion Krause-Wolters (University of Cologne), Valerie Lemke (University of Cologne),Marije Lesterhuys (University of Antwerp), Stefan Martens (University of Antwerp), Prof. dr. Nele Michels (University of Antwperp), Filip Moens (AHOVOKS), Michèle Pettinato (CLiPS University of Antwerp), Prof. dr. Gert Rijlaarsdam (University of Amsterdam), Iris Roose (Potential Project), dr. Pierpaolo Settembri (College of Europe), Prof. dr. Jean-Michel Rigo (Hasselt University), dr. Joke Spildooren (Hasselt University), dr. Sabine Stephany (University of Cologne), dr. Olia E. Tsivitanidou (University of Cyprus), Andries Valcke (Headmaster Training Flemish Public Schools), Danielle Van Ast (Flemish Public Schools Antwerp), Tine van Daal (University of Antwerp), Marie-Thérèse van de Kamp (University of Amsterdam), Kirsten Vandermeulen (Thomas Moore University of Applied Sciences), Roos Van Gasse (University of Antwerp), Prof. dr. Hilde Van Keer (Ghent University), Kristof Vermeiren, Ellen Volckaert (Hudson), Prof. dr. Jo Verhoeven (University of Antwerp) and, Ivan Waumans (Karel de Grote University College).

Data availability statement

The data and R script that support the findings of this study are openly available in the Zenodo repository at http://doi.org/10.5281/zenodo.1493425

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is part of a larger project (D-PAC) funded by the Flanders Innovation & Entrepreneurship and the Research Foundation (grant number 130043).

Notes on contributors

San Verhavert is working on the Digital Platform for the Assessment of Competencies project (D-PAC) at the University of Antwerp (Belgium). His PhD focuses on the method of comparative judgement.

Renske Bouwer is now assistant professor in Pedagogical and Education Sciences at the Vrije Universiteit Amsterdam. At the time of this research she was research coordinator for the D-PAC project at the University of Antwerp. Her own research focuses on the quality of comparative judgement for the assessment of writing quality and the effects of comparative judgements on student's learning.

Vincent Donche is an associate professor in Training and Education Sciences at the University of Antwerp, Belgium. His research interests are situated in the domains of student learning, higher education, assessment and related educational measurement issues.

Sven De Maeyer is a full professor in Training and Education Sciences at the University of Antwerp. He has expertise in statistical modelling. His research mainly focusses on assessment in both education and vocational contexts, with a strong focus on judgement and rater-effects and the merits and pitfalls of comparative judgement.

ORCID

San Verhavert http://orcid.org/0000-0003-0633-9753
Vincent Donche http://orcid.org/0000-0002-9405-3896

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*, 371–383.

Bi, J. (2003). Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies*, *18*, 61–76.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*, 324–345.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London, U.K.: Qualifications and Curriculum Authority.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment* (Cambridge Assessment research report). Retrieved from www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives*, *25*(2), 1–24.

Bramley, T, & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*, 43–58. doi:10.1080/0969594X.2017.1418734

Cho, K., & Schunn, C. D. (2018). Finding an optimal balance between agreement and performance in an online reciprocal peer evaluation system. *Studies in Educational Evaluation*, *56*, 94–101.

Core Team, R. (2017). R: A language and environment for statistical computing (Version 3.4.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Dochy, F. J. R. C., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation*, *23*, 279–298.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111–127.

Goossens, M., Bouwer, R., & De Maeyer, S. (2017). The reliability and validity of peer assessment based on comparative judgements. Presented at the Assessment in Higher Education, Manchester, UK.
Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–19.
Humphry, S., & Mcgrane, J. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, *42*, 443–460.
Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*, 1774–1787.
Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, *47*, 93–101.
Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.
Kaslow, N. J., Rubin, N. J., Bebeau, M. J., Leigh, I. W., Lichtenberg, J. W., Nelson, P. D., … Leon, I. (2007). Guiding principles and recommendations for the assessment of competence. *Professional Psychology: Research and Practice*, *38*, 441–451.
Kimbell, R. (2007). E-assessment in project e-scape. *Design and Technology Education: an International Journal*, *12*, 2. Retrieved from https://ojs.lboro.ac.uk/DATE/article/view/Journal_12.2_0707_RES6
Laming, D. (2003). *Human judgment: The eye of the beholder* (1st ed.). London: Cengage Learning EMEA.
Lane, S., & Stone, C. A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement*, (pp. 387–432). Westport, CT: American Council on Education & Praeger.
Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality: An assessor's perspective* (Doctoral dissertation). University of Antwerp, Antwerp, Belgium.
Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., & De Maeyer, S. (2018). When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality. *L1 Educational Studies in Language and Literature*, *18*. doi:10.17239/L1ESLL-2018.18.01.02
Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*, 553–568.
Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.
McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, *22*, 368–389.
McMullan, M., Endacott, R., Gray, M. A., Jasper, M., Miller, C. M. L., Scholes, J., & Webb, C. (2003). Portfolios and assessment of competence: A review of the literature. *Journal of Advanced Nursing*, *41*, 283–294.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
Pollitt, A., & Murray, N. L. (1995). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 74–91). Cambridge, U.K.: Cambridge University Press.
Pollitt, A. (2004). Let's stop marking exams. Presented at the IAEA Conference, Philadelphia, PA.
Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, *19*, 281–300.
Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 1–19). Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-8905-3_4
Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*, 211–223.
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability". *Practical Assessment, Research & Evaluation*, *9*, 1–11.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.
Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2014). The standard error of measurement is a more appropriate measure of quality for postgraduate medical

assessments than is reliability: An analysis of MRCP(UK) examinations. *BMC Medical Education*, *10*, 1–9.
van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*. doi:10.1080/0969594X.2016.1253542
van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M.-T., Donche, V., & De Maeyer, S. (2017). The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. *Frontiers in Education*, *2*.
Van Gasse, R., Mortier, A., Goossens, M., Vanhoof, J., Van Petegem, P., Vlerick, P., & De Maeyer, S. (2017). Feedback opportunities of comparative judgement: An overview of possible features and acceptance at different user levels. In D. Joosten-Ten Brinke & M. Laanpere (Eds.), *Communications in Computer and Information Science* (Vol. 653, pp. 23–38). Cham, Switzerland: Springer.
Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2018). a meta-analysis on the reliability of comparative judgement [dataset], Zenodo. doi:10.5281/zenodo.2586084
Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*, 428–445.
Wheadon, C. (2015, September 22). The opposite of adaptivity [Blog post]. Retrieved March 15, 2017, from blog.nomoremarking.com/how-many-judgements-do-you-need-cf4822d3919f
Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement* (1st ed. ed.). Chicago, IL: MESA Press.

Appendix A. Table with Correlations Between the Variables

Table A1. Correlation between variables.

Variable	SSR [a]	N_C [b]	N_R [c]	N_{CR} [d]	N_A [e]	N_{RA} [f]
SSR [a]	–	.11	−.10	.51	.01	−.07
N_C [b]	.11	–	.94	.19	.38	.28
N_R [c]	−.10	.94	–	−.12	.38	.35
N_{CR} [d]	.51	.19	−.12	–	−.03	−.15
N_A [e]	.01	.38	.38	−.03	–	−.26
N_{RA} [f]	−.07	.28	.35	−.15	−.26	–

Note
[a] SSR = Scale Separation Reliability
[b] N_C = Number of comparisons per representation
[c] N_R = Number of representations
[d] N_{CR} = Number of comparisons per representation
[e] N_A = Number of assessors
[f] N_{RA} = Number of representation per assessor